Обучение машинного перевода без параллельных текстов

Борисова А. В. Бахтеев*[1] *О. Ю. Стрижов*[2] *В. В.
[1]Московский физико-технический институт
[2]Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН

В данной работе рассматривается задача построения системы машинного перевода. Для построения систем машинного перевода используются параллельные предложения. Однако не для всех пар языков существует достаточный массив размеченных предложений. В данной работе рассматривается метод построения модели машинного перевода без использования параллельного корпуса. Модель основана на Seq2Seq, предложениям обоих языков ставится в соответствие вектор в общем скрытом пространстве. Вычислительный эксперимент проводится на паре языков «русский-украинский».

Ключевые слова: *машинный перевод, Seq2Seq.*

1 Введение

Использование глубоких нейронных сетей хорошо показывает себя в задачах машинного перевода с параллельными предложениями [1,2]. Однако в таких моделях для высокого качества перевода критичен размер выборки. Необходимо несколько миллионов пар предложений.

Для некоторых пар языков не существует достаточного количества размеченных предложений. Было предложено несколько подходов к решению этой проблемы [3, 4]. Одним из подходов является использование результатов предыдущих итераций нейронной сети для пополнения выборки [5].

В [6] для перевода между английским и французским языками используется модель Seq2Seq. В модели используются кодировщики, рекуррентные нейронные сети LSTM, переводящие предложение в скрытое векторное пространство, и дискриминатор, определяющий язык по векторному представлению предложения. Кодировщики оптимизируются таким образом, чтобы дискриминатор не мог определить язык предложения по его скрытому представлению. Как результат, скрытые представления одного и того же предложения на разных языках совпадают. Для уменьшения переобучения в предложения добавляют случайный шум.

Шаг обучения нейронной сети состоит из двух фаз. На первой оптимизируются параметры сети-дискриминатора. На второй стадии происходит оптимизация сетей-кодировщиков.

В данной работе рассматривается подобный подход для перевода в паре языков "русский-украинский". Качество перевода оценивается с помощью метрики BLEU.

2 Постановка задачи

Пусть заданы выборки:
$\mathfrak{D}^{src} = \{\mathbf{s}_1^{src}, ..., \mathbf{s}_{m_{src}}^{src}\}$ - обучающая выборка на первом языке,
$\mathfrak{D}^{tgt} = \{\mathbf{s}_1^{tgt}, ..., \mathbf{s}_{m_{tgt}}^{tgt}\}$ - обучающая выборка на втором языке.
Здесь $\mathbf{s}_i^{src}$ - предложение на первом языке, $\mathbf{s}_i^{tgt}$ - предложение на втором языке.
Введем обозначения для мощности словарей: V^{src} и V^{tgt}.
Наборы $\mathfrak{D}^{src}$ и $\mathfrak{D}^{tgt}$ могут не являться параллельными.

$\mathfrak{D}^{valid} = \{(\mathbf{s}_1^{src}, \mathbf{s}_1^{tgt}), ..., (\mathbf{s}_{m_{valid}}^{src}, \mathbf{s}_{m_{valid}}^{tgt})\}$ - валидационная выборка, представляющая из себя корпус параллельных предложений.

Под $\mathbf{s}(k)$ будем понимать k-е слово прделожения $\mathbf{s}$. $d(.,.)$ - метрика между словами.

В задаче минимизируется функция потерь:

$$L = \sum_{i=1}^{m_{valid}} \sum_{k=1}^{\mathbf{s}_{m_{valid}}^{tgt}} d(\mathbf{g}^{tgt}(\mathbf{f}^{src}(\mathbf{s}_i^{src}))(k), \mathbf{s}_i^{tgt}(k))$$

В модели используются кодировщик $\mathbf{f}$ и декодировщик $\mathbf{g}$ - рекуррентные нейронные сети. Кодировщик отвечает за перевод предложений обоих языков в скрытое пространство. Декодировщик - за обратное отображение из скрытого пространства в предложения двух языков. Перевод предложений разных языков использует разные словари. Обозначим это при помощи индексов: $\mathbf{f}^{src}$ и $\mathbf{g}^{src}$ - кодировщик первого языка, $\mathbf{f}^{tgt}$ и $\mathbf{g}^{tgt}$ - кодировщик второго языка.

Для обучения не используются параллельные предложения. Основной идеей обучения модели будет нахождение общего скрытого пространства для двух языков. Для этого используется сеть дискриминатор, которая по представлению h предложения в скрытом пространстве определяет язык исходного предложения. Дискриминатор оптимизируется с целью наилучшего распознавания языка, кодировщик - с целью генерации наиболее похожих представлений для разных языков.

Запишем функцию потерь.Для избежания переобучения перед каждым кодированием к предложению добавляется шум $\sigma(x)$.На первом шаге будем оптимизироваться функция:

$$L_{AE} = \sum_{i=1}^{m_{valid}} \|\mathbf{g}(\mathbf{f}(\sigma(\mathbf{s_i}))) - \mathbf{s_i}\|$$

Следующий шаг использует модель слабого перевода $\hat{\mathbf{g}}$. Функция потерь имеет вид:

$$L_{TR} = \sum_{i=1}^{m_{valid}} \|\mathbf{g}(\mathbf{f}(\hat{\mathbf{g}}(\mathbf{f}(\mathbf{s_i}))) - \mathbf{s_i}\|$$

Последний шаг - оптимизация дискриминатора таким образом, чтобы он различал векторные представления разных языков в скрытом пространстве:

$$L_{ADV} = \sum_{i=1}^{m_{valid}} (\log p(language = src|f(x_i)) + \log p(language = tgt|f(y_i)))$$

Итоговая функция оптимизации:

$$L = (L_{AE}, L_{TR}, L_{ADV}) * \boldsymbol{\omega} \to \min$$

$\boldsymbol{\omega}$ - вектор искомых параметров.

3 Базовый алгоритм

3.1 Получение слабого перевода

Сгенерирован словарь пар слов на основе смежных слов в двух парах словарей "русско-английский"и "англо-украинский"из [7]. Данные словари можно считать реальными выборками. Далее строится алгоритм, который делит предложения на слова и для каждого

ищет значения в сгенерированном словаре. Если значения не находятся, возвращается оригинальное слово.

Оценка работы алгоритма проводится с помощью BLEU-метрики. Результаты работы базового алгоритма на реальной выборке из субтитров к одному фильму на паре языков "русский-украинский"оценен:

$$BLEU = 10.86, 27.2/12.9/7.8/5.1(BP = 1.000, ratio = 1.010, hyp\_len = 3308640, ref\_len = 3275742)$$

3.2 Альтернативный метод

Берутся два словаря из [7], соотносящие каждому слову некоторое векторное представление, причем векторное пространство общее и синтетически размечено так, что для слов, являющимися реальными переводами друг друга в разных языках, векторное представление приблизительно одинаковое. Мы строим алгоритм, переводящий слово в векторное пространство и в нем ищем наиболее близкие векторные представления из другого языка. Расстояние оценивается по L2 норме.

Литература

[1] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016.

[2] KyungHyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *CoRR*, abs/1409.1259, 2014.

[3] Will Y. Zou, Richard Socher, Daniel Cer, and Christopher D. Manning. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398. Association for Computational Linguistics, 2013.

[4] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc., 2014.

[5] Nicola Bertoldi and Marcello Federico. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, StatMT '09, pages 182–189, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[6] Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. *CoRR*, abs/1711.00043, 2017.

[7] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017.

Поступила в редакцию